Journal of University of Babylon for Pure and Applied Sciences, Vol. (27), No. (1): 2019

Writer Identification Based on Arabic Handwriting Recognition by using Speed up Robust Feature and K- Nearest Neighbor Classification

[1] Alia Karim Abdul Hassan [2] Bashar Saadoon Mahdi [3] Asmaa Abdullah Mohammed
[1,2,3] *Computer Science Department, University of Technology, Baghdad, Iraq*

[1] hassanalia2000@yahoo.com, [2] basharsadoon@yahoo.com [3] asmaaabdulah848@gmail.com,

ARTICLE INFO
Submission date: 13/8/2018
Acceptance date:15/10/2018
Publication date: 10/1/2019

Abstract

In a writer recognition system, the system performs a "one-to-many" search in a large database with handwriting samples of known authors and returns a possible candidate list. This paper proposes method for writer identification handwritten Arabic word without segmentation to sub letters based on feature extraction speed up robust feature transform (SURF) and K nearest neighbor classification (KNN) to enhance the writer's identification accuracy. After feature extraction, it can be cluster by K-means algorithm to standardize the number of features. The feature extraction and feature clustering called to gather Bag of Word (BOW); it converts arbitrary number of image feature to uniform length feature vector. The proposed method experimented using (IFN/ENIT) database. The recognition rate of experiment result is (96.666).

Keywords: IFN/ENIT Database; SURF feature extraction; K-mean algorithm; KNN classifier algorithm.

1. Introduction

Handwriting has always played considerable role in people's lives. Even after the invention of innovative smart devices (such as, I-Pads, smart phones, and so on), people still have a preference for writing. Therefore, the number of hand-written documents around us keeps increasing daily. Writer recognition is originated from the older and broader automatic handwriting identification of domain writer recognition users strive for actually the opposite with aiming at maximally exposing the details of individual hand-writing styles for writer identification. It should be mentioned that writer recognition could get rid of particular ambiguities in pattern identification procedure. In the mode of recognition, the system takes an unknown example as an input and has the task of identifying the author from a group of writers which are stored and known by the system [1]**.** Most conducted studies in writer identity recognition and confirmation have been offline and signature-based online methods are more common in identity recognition. Most studies in identity recognition focus on the English language and there has been few studies regarding Arabic and Persian handwritten texts in comparison to the English language [2]**.**

Systems of writer identification are concerned with the comparison of unknown written documents (queries) to known written ones (i.e. samples) for the aim of determining whether the two documents are written by the same person. The system of writer recognition is involved with two fundamental categories: distinct variations of hand-writing that are distinct features and the individuals that are class properties. Characteristics for the discrimination of styles of hand-writing play a significant part in systems of writer recognition [3]. Hand-writing identification is the capability of a device in receiving and interpreting readable hand-written input from sources like printed documents, photos, touchscreens, etc. The image of the written text could be captured "offline" from a paper using optical scanners (using the Optical Character Recognition (OCR) technology) or intelligent word identification. On the other side, the movements of the pen could be captured "online", via a pen-based computer screen surface for example [4]**.**

The researches for writer recognition based on Arabic handwriting word identifying are very limited in comparison with the Latin texts. Mohamed N. Abdi [5] proposed a system for "Arabic Writer Identification and Verification using Template Matching Analysis of Texture"**.** Intensive tests have been performed with 557 hand-writing instances from a hundred of different writers in the IFN/ENIT data-base, and this research showed good rates for writer recognition that reached 85% for Top1, 90% for Top2 and 95% for Top10. And concerning the task of verification, a relatively good equal error rate (EER) of 5.9% ± 0.11% has been reached. Hannad, Y., Siddiqi, I., & El Kettani, M. E. Y [6] proposed a system for Writer recognition by using texture descriptors of handwritten fractions. Texture descriptors, including histograms of Local Binary Patterns (LBP), Local Ternary Patterns (LTP) and Local Phase Quantization (LPQ) are then computed from these fractions. The proposed system used two databases, one in Arabic and one for English in Arabic database IFN/ENIT the result for writer identification rate of 94.89. Hannad, Y., Siddiqi, I., El Merabet, Y., & El Youssfi El Kettani, M [7] proposed a system for Writer Recognition System by Using the Histogram of Oriented Gradients (HOG) of Handwritten. In proposed system using IFN/ENIT database on 411 writers, the system result identification rate of 86.62% .Sheikh, A., & Khotanlou, H [8] have suggested a system for Writer Identity Recognition particularly at the feature extraction stage (since, manual feature extraction has taken long time) this study included two basic steps: the first was dedicated to training and the second concerned testing. For assessing the efficiency of the introduced features, the Hidden Markov Model (HMM) has been utilized as the classifier. Increasing the accuracy and credibility of the identity recognition system due to using the chain network at the training stage.

2. Speed up Robust Features (SURF)

SURF is a robust technique for object recognition, image registration and matching (first presented by Herbert Bay in 2006). This algorithm is based on the same ideas and steps the SIFT was based on, with the difference that it uses a different scheme and it should provide more efficient and less time consuming results. This method uses detection, description and matching. For the detection of interest points, SURF utilizes determinant of Hess blob detector [9].

2.1 Detecting interest points

The approach that has been used in this paper for the detection of interest points utilizes quite a basic Hess matrix approximation. Which lends itself to using integral images to be made common by "Viola and Jones", which greatly reduced the time of

computation. Integral images fit in the more generalized model of box lets, as suggested by "Simard".

2.1.1 Integral images

For the sake of making the research more self-contained, it includes a brief discussion of the idea of integral images. They guarantee fast calculation of box type convolution filters. The entry of an integral image at $I\sum(X)$a location x=(x, y) depicts the summation of all pixels in the input image I in a rectangular area produced by the origin and x.

$$I\sum(x) = \sum_{i=0}^{i\leq x}\sum_{j=0}^{j\leq y} I(i.j) \quad (1)$$

As soon as the integral image is obtained, it takes 3 addition operations for the calculation of the summation of the intensities over any upright, rectangular region (Figure 1). Therefore, the time of calculation is invariant to its size. Which is an important element in this approach, due to the fact that big filter sizes are used in this research.

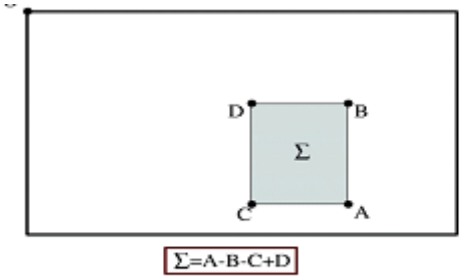

Figure 1: Integral images

2.1.2 Hess matrix-based points of interest

Hess matrix is the foundation of SURF. The local maximum of its determinant can determine the position and scale of feature points. SURF obtains stable points by the use of the Hess matrix to find the extrema points, and uses the maximal value of the matrix determinant to mark the position of the blob-like structure. The determinant of a Hess matrix reflects the extent of the response and is an expression of the local variation that surrounds the region, as shown in the equations**.**

$$H(X.\sigma) = \begin{bmatrix} Lxx(x.\sigma) & Lxy(x.\sigma) \\ Lxy(x.\sigma) & Lyy(x.\sigma) \end{bmatrix} \quad (2)$$

Where $Lxx(x, \sigma)$ denotes the convolution of the Gauss 2nd derivative $\frac{\sigma 2}{\sigma x 2} g(\sigma)$ with the image I in point X. and the same way for Lxy(x, σ)and Lyy(x, σ).

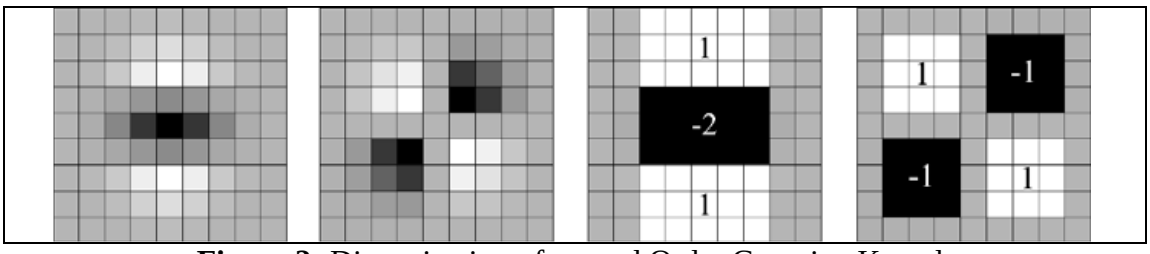

Figure 2: Discretization of second Order Gaussian Kernels

$Lxx(x,\sigma)$is the convolution of the image with the second derivative of Gaussian. The heart of the SURF detection is non-maximal suppression of the determinants of the hessian matrices. It is costly to calculate convolution, so it is approximated and speeded up with the use of integral images and approximated kernels. The second order Gaussian kernels $\frac{\partial^2}{\partial y^2} g(\sigma)$ must be discretized and cropped before applying them, rectangular boxes, gray regions, corresponds to 0 in the kernel while white are positive and black are negative. By this way it is possible to calculate the approximated convolution by Equation (3). The approximated and discrete kernels are corresponding to as D_{yy} for L_{yy} (x, σ) and D_{xy} for L_{xy} (x, σ), w is weight to assure the energy conservation for the Gaussians. The w term is theoretically sensitive to scale but it can be keptconstant at 0.9 [9].

$$Det(H_{approximation}) = D_{xx}D_{yy} - (wD_{xy})^2 \qquad (3)$$

2.1.3 Scale space representation

Points of interest must be obtained at various scales, not least due to the fact that the searching for correspondences typically needs them to be compared in images where they're viewed at various scales. Scale spaces are typically applied as a pyramid of images. Which are repeatedly smoothed using Gauss blurring and afterwards, those smoothed images are sub-sampled for the aim of achieving a higher level in the pyramid. Then, the layers of this pyramid are subtracted to get the DoG images in which edges and blobs could be obtained. The scale space is split to octaves. Each one of which represents a set of filter response maps that have been obtained via performing a convolution of the same input image with a filter of increasing size. Generally, an octave encompasses a scaling factor of 2.

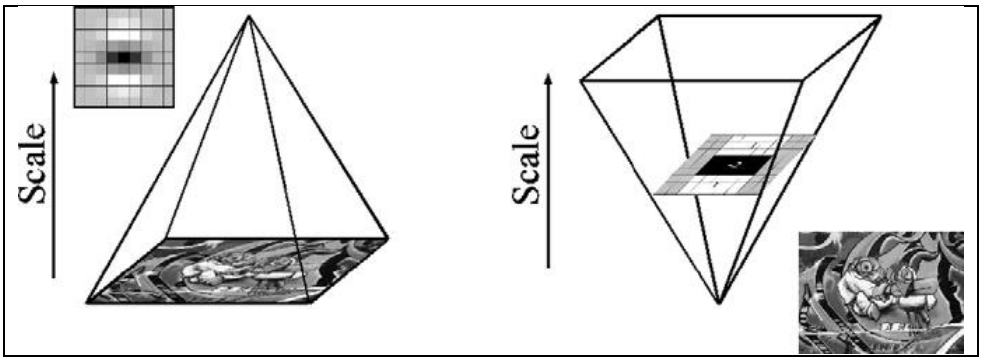

Figure (3): iterative reduction of the image size (left), up-scaling the filter at constant cost (right).

2.1.4 Key point description

Descriptor provides an individual and a robust description of the obtained property, it is possible to generate it according to the adjacent region of a point of interest, and descriptor may be computed efficiently with integral images due to the fact that SURF depends on the responses of Haar wavelet. SURF has to decide the orientation in order to be capable of achieving the invariance to rotations, the area of interest is split to 4x4 sub regions which denoted by the values of a wavelet response in the directions of x and y, the wavelets response in the x and y direction is known as dx and dy.For every sub region, a vector v is computed, depended on 5x5 samples as show in figure (4). The descriptor of the points of interest is that the 16 vectors for the sub region concatenated as equation [9].

$$V = \{\sum dx. \sum |dx|. \sum dy. \sum |dy|\} \quad (4)$$

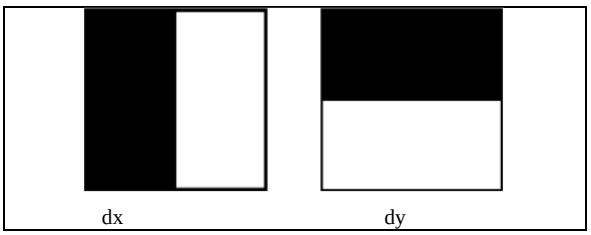

Figure (4): wavelet response W=-1for black area and w=1 for weight area for Haar kernals

3. K-Mean Clustering: -this is an approach of clustering observations to a certain number of distinct clusters. "K" denotes the number of given clusters. There are different distance measurements for determining what observation to append to what cluster. This approach has the aim of minimizing the measure between the cluster centroid and the specified observation via the iterative appending of an observation to any of the clusters and end the process after achieving the minimal distance measure.

3.1 The Steps of K-Mean Algorithm [10]

Step1-Get vector of detected key points number of clusters K.

step2-Find centroid for key points.

step3-Calculate distance between keypoints and centroid.

step4-Grouping keypoints according to minimum distance.

step5-Repeat steps 2,3 and 4 to the points where the centroids no longer move output clustered keypoints.

4. Bag of word (BOW)

The feature extraction and feature clustering by K-means cluster called to gather Bag of Word (BOW), it converts arbitrary number of image feature to uniform length feature vector. One of the most generalized and commonly utilized approaches for class identification, it is also referred to as bag of features or bag of key points model. This approach produces a histogram, which includes the distribution of visual words that are obtained from the testing image, and after that, classifiers perform the classification of the image according to the features of every classifier, the K-NN classifier utilizes the testing image histogram and a learned architecture from the set of training for the prediction of a class for the testing image. The purposed model of Bow is the representation. It deals with detecting features and image representation. Features have to be obtained from images for the sake of representing the images in the form of histograms. We extracted features using SURF. The reason for using BoW Model is the image representation. The features that have been detected by SURF descriptors have been put into a code-book by k-means clustering. Now, the classification may be implemented via performing a comparison of the histograms that represent the code words [11].

5. Knearst Neighbor Algorithms

It is an approach that is utilized in pattern recognition for classification; it is used for prediction or regression as well. It's a non-parametric approach because the data have no statistical properties, a supervised learning because it considers the training prototypes for the final decision, and could also be referred to as lazy learning because the generalization isn't performed prior to the query point or instance based learning (it performs a comparison of the query point with data for training). Even in the case where there isn't any explicit training process, it's known as supervised learning, due to the fact that the classification is determined based on the prototypes (a group of data with previously known classes) [12]. This approach is popular in the area of pattern identification for classifying a pattern according to the nearest training instances in the property space. K-NN is considered one of the simplest machine leaning approaches [13].

KNN is an approach for object classification on the basis of the nearest training samples in the property space. K-NN is one of the simplest machine learning methods. Training procedure for this method only includes storing property vectors and labels of the training images. In the procedure of classification, the unlabeled query point is appointed to the label of its k nearest neighbors. Usually, the object is classified according to the labels of its k-NNs via majority vote. For example, in Figure 5, we notice KNN classification method. If K=3, test sample is classified into class1, due to the fact that there are two triangles and only one square within the inner circle. The final voting (decision) is class1(Red triangle).

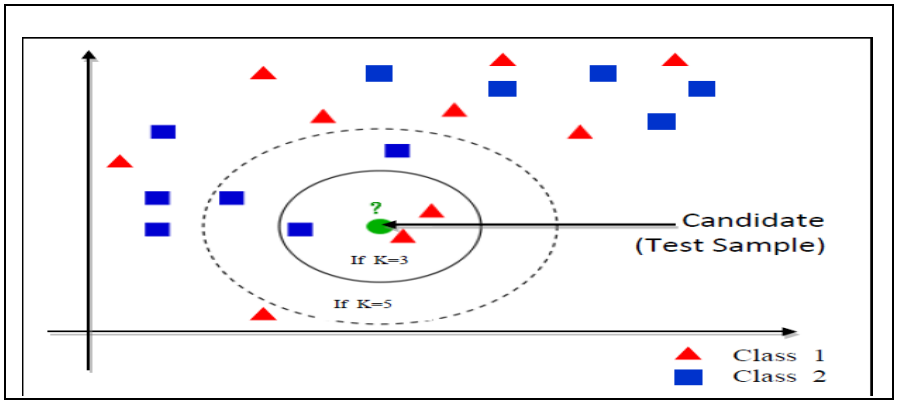

Figure (5): An example of KNN Technique in classification

If K=5, test sample is classified to class2= square (Blue). There are 3 squares and only 2 triangles inside the inner and outer circles. The final voting is class2 (Blue square). Euclidean distance function has been used [13]:

$$D(x.y) = \sqrt{\sum_{k=1}^{N}(X_k - y_k)^{\ 2}} \qquad (5)$$

Where D is the x length, x_k is the first vector, and y_k is the second vector. A main advantage of the K-NN algorithm lies in the fact that it's very easy to be implemented. As it has been previously stated, it's a lazy learning method and this, does not require any training before making real time predictions. Which makes K-NN considerably faster than other algorithms requiring training such as SVMs, regression, due to the fact that the algorithm does not require any training

prior to making predictions, new data may seamlessly be added. There are only 2 parameters that are required for implementing KNN, the value of K and the distance function (Euclidean or Manhattan for example). The main drawback of the K-NN lies in the fact that it utilizes all the characteristics equally in computing for similarities. Which could produce classification errors, especially when there's only a small sub-set of properties useful for classifying [11].

6. Proposed Writer Identification System

Figure 6 shows the major stages of the proposed method for writer Identification based on Arabic hand-writing word recognition. Algorithem-1 illustrate the proposed system steps. In Algorithem-1 **the first** input Arabic handwriting word image**s. Step2** preprocessing image most of the identifying and classification techniques require the data to be in a predefined type in this model we convert the input images to gray image to easy implementation in next step.

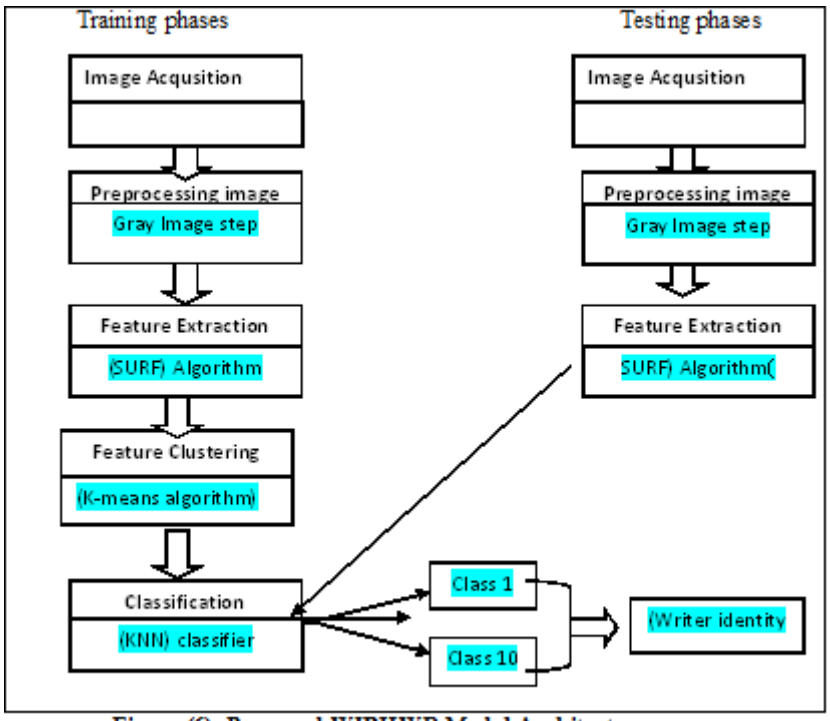

Figure (6): Proposed WIBHWR Model Architecture

In step3 feature extraction is a very significant step in systems of word identification and for numerous pattern identification tasks. It has the aim to delete redundancy from the data and provide a more efficient representation of the text image we used speed up robust feature (SURF) algorithm . For example in figure 7.

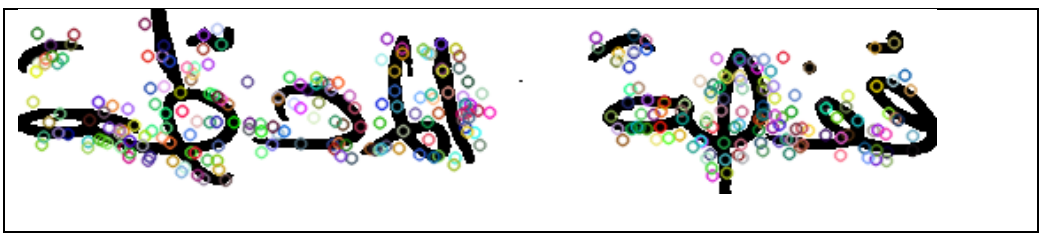

Figure (7): detect features in (IFN/ENIT) database by surf algorithm

Fourth step due to the fact that the detected keypoints extracted by SURF are different in count between one image to the other, so in this paper utilizes K-means in order to make cluster of the detected key points be on the same counter. After the extraction of properties from each of the testing and training images we also used for that set of feature to cluster it into groups in order to achieve this, k-means clustering has been performed over all of the vectors. K-means is an approach for clustering or dividing n observations or, in the case of this paper, features to k clusters where every one of the features belongs to the cluster of its closest average. We have clustered the features and prepared data for Histogram Generation, where every patch with an image is mapped into a specific code-word via k-means clustering procedure and therefore, every image may be denoted by a code word histogram. For the match with which feature clusters in the database. This is the final step which is preceding to the classification itself. While the **fifth step: -**is the writer identification after the features clustering then input in the K Nearest Neighbor (KNN) works by classifying the whole handwritten word into its writer. Each handwritten is classified into its desired class, each with a class label as training samples which are stored at the training phase. In the classification phase, the distances (the Euclidean distance is more popular) between each training sample and tested sample is calculated. K is a user-defined constant, The *K* training samples that have the smallest distances (nearest) to the test sample are found and recognition their labels. By using the majority vote on the neighbor samples will have declared the class of the test sample. **Output:** the output of the proposed system is class label of desired writer; each class label of the input word is assigned to its writer to make the output of the proposed system display the writer name.

Algorithm-1: Proposed *writer identification based on Arabic handwriting, System*
Input: *word image**s** for each writer in dataset**.***
Output: *class label of desired writer*

Step1: *Preprocessing input image and convert color image(RGB) to gray image.*
Step2: *To delete the redundancy of data and obtain a better representation of the text image, apply SURF algorithm on Arabic handwriting word images to extract feature.*
Step3*: cluster extracted features into groups using k-means. Use clustered properties for generating the histogram. Every cluster of the images can be represented with a histogram of the code words.*
Step4*: Apply K nearest neighbor (KNN) to classifier the Arabic handwriting word image to true writer.*
Step 5*:* ***return(class label)***

7. Experimental Results

The experimental perform on personal computer that has properties The suggested approach is implemented with the use of visual C++ R2013a version, under windows-7 64-bit OS, with RAM 4GB, CPU 2.50GHz core i5 and proved to be fast and efficient. This work is experimented on dataset (IFN/ENIT) which consists of 32492 word images. we take 70 for the training data and 30 for testing data and we in the first convert the images to gray image in preprocessing step, then feature extraction by using surf algorithm then to find the clusters and center and order without repeating we use (K-Means) algorithms then the result input in the K-nearest

neighbor algorithms (KNN) to classifier. The proposed system compared with other related works using the traditional evaluation measure (precision) as shown in table-1.

Table1: Recognition results of various systems

Author	Years	Techniques	Precision
[5]	2012	Template Matching Analysis of Texture	85% for Top1, 90% for Top2 and 95% for Top10
[6]	2016	(LBP)+ (LTP)+ (LPQ)	94.89%
[7]	2016	HOG	86.62%
The proposed system	2018	SURF+KNN	96.666

8. Conclusion

The proposed system identifying the handwritten Arabic word as one entity without segmentation to sub letters. Feature extraction is very valuable stage in word identification systems and for a wide range of pattern identification tasks. It has the aim of removing redundancy from the data and reach a better representation of the text image with a group of numerical features using (SURF) algorithm to feature extraction, used K-means algorithm to feature clustering and the both feature extraction and feature clustering called Bag of word (BOW) .in this paper, we used (KNN) which is one of the soft computing techniques for classification the handwritten word image into its writer. It has been noticed that the rate of success of any system of identification is not merely dependent on the extraction of features but it is also dependent on numerous reasons like the preprocessing step and the classifier, recognizer technique. We used the dataset (IFN/ENIT). Result precision is 96.666%.

9. Future Work

Enhancing the proposed system to be able of working in real time via applying it in on-line identification of Mobile and tablet devices. The suggested system can be enhanced to operate with printed Arabic document identification. The plan is extending this study to various feature extraction and classification methods to minimize error rates.

Conflict of Interests.
There are non-conflicts of interest .

Reference

[1] A. Bensefia and T. Paquet, "Writer verification based on a single handwriting word samples," *EURASIP J. Image Video Process.*, vol. 2016, no. 1, p. 34, 2016.

[2] H. Khotanlou, "Writer Identity Recognition and Confirmation Using Persian Handwritten Texts," *Adv. Comput. Sci. an Int. J.*, vol. 4, no. 6, pp. 24–30, 2015.

[3] J. H. Y. AlKhateeb, "Word based off-line handwritten Arabic classification and recognition. Design of automatic recognition system for large vocabulary offline handwritten Arabic words using machine learning approaches." University of Bradford, 2010.

[4] P. S. A. Chakravarthy, A. S. N., Penmetsa V. Krishna Raja, "Handwritten Text Image Authentication Using Back Propagation," *arXiv Prepr. arXiv*, no. 1110.1488, 2011.

[5] M. N. Abdi and M. Khemakhem, "Arabic writer identification and verification using template matching analysis of texture," *Proc. - 2012 IEEE 12th Int. Conf. Comput. Inf. Technol. CIT 2012*, no. December, pp. 592–597, 2012.

[6] Y. Hannad, I. Siddiqi, and M. E. Y. El Kettani, "Writer identification using texture descriptors of handwritten fragments," *Expert Syst. Appl.*, vol. 47, pp. 14–22, 2016.

[7] Y. Hannad, I. Siddiqi, Y. El Merabet, and M. El Youssfi El Kettani, "Arabic writer identification system using the histogram of oriented gradients (HOG) of handwritten fragments," in *Proceedings of the Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, 2016, pp. 98–102.

[8] H. Sheikh, A., Khotanlou, "Writer identity recognition and confirmation using persian handwritten texts," *Int. J. Adv. Appl. Sci.*, vol. 6, no. 2, pp. 98–105, 2017.

[9] L. Bay, H., Ess, A., Tuytelaars, T., & Van Gool, "Speeded-up robust features (SURF).," *Comput. Vis. image Underst.*, vol. 3, no. 110, pp. 346–359, 2008.

[10] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J.*, vol. 1, no. 6, pp. 90–95, 2013.

[11] J. Kim[1], B. S. Kim, and S. Savarese, "Comparing image classification methods: K-nearest-neighbor and support-vector-machines," in *Proceedings of the 6th WSEAS international conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics*, 2012, vol. 1001, pp. 42122–48109.

[12] V. L. Boiculese, G. Dimitriu, and M. Moscalu, "Improving recall of k-nearest neighbor algorithm for classes of uneven size," in *2013 E-Health and Bioengineering Conference (EHB)*, 2013, pp. 1–4.

[13] J. H. AlKhateeb, F. Khelifi, J. Jiang, and S. S. Ipson, "A new approach for off-line handwritten Arabic word recognition using KNN classifier," in *2009 IEEE International Conference on Signal and Image Processing Applications*, 2009, pp. 191–194.

[14] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Nat. Lang. Eng.*, vol. 16, no. 1, pp. 100–103, 2010.